Proc. of the First International Conference on Intelligent Interactive Technologies and Multimedia, Dec. 2010, Allahabad, India.

Pronunciation scoring for Indian English learners using a phone recognition system

Chitralekha Bhat, K.L. Srinivas, Preeti Rao
Department of Electrical Engineering,
Indian Institute of Technology Bombay,Mumbai 400076,India
{chitragb,k.l.srinivas,prao}@ee.iitb.ac.in

ABSTRACT

Feedback on pronunciation or articulation is an important component of spoken language teaching. Automating this aspect with speech recognition technology has been an active area of research in the context of computer-aided language-learning systems. Well-known limitations in the accuracy of automatic speech recognition (ASR) systems pose challenges to the reliable detection of pronunciation errors in the speech of non-native speakers. We present the design of a pronunciation scoring system using a phone recognizer developed with the popular HTK and CMU Sphinx HMM-based ASR toolkits. The system is evaluated on Indian English speech in the realistic situation where there is no matching database available for training the speech recognizer. Different approaches to the training of acoustic models and to constraining the phone recognition system are investigated.

1. INTRODUCTION

Fluency in speech by a non-native speaker of a language can be judged based on the correctness of pronunciation and prosody. Feedback on pronunciation or articulation is an important component of spoken language teaching. Non-native language learners can differ prominently from native speakers with respect to speech articulation as well as prosody. Automatic detection of specific speaking errors can provide for valuable feedback for the learner. Such a facility requires the reliable segmentation of the non-native speaker's utterance at the phone level including detection of disfluencies and the extraction of prosodic attributes such as duration and pitch at the syllable level that can then be compared with the corresponding attributes of native speech. This work deals only with the detection of articulation or pronunciation errors of read speech. Non-native speakers tend to mispronounce words by substituting phones from their native language and also make phone insertion and deletion errors, influenced by phonotactic constraints of their own native language (L1).

ASR technology would seem to provide the solution to automatic pronunciation error detection by the ability to decode speech into word and phone sequences and provide acoustic likelihood scores indicating the match with trained native speech models. However, state-of-the-art ASR systems fare poorly on phone recognition accuracy unless aided by powerful language models. Further, the phone error rates rise steeply when there is mismatch between test and train data as would be expected with non-native speakers due to accent variations from the native speech upon which the recognizer has been trained. Access to a non-native speech database in the target language (L2) can help to reduce the mismatch via better trained acoustic models. However due to the typical non-homogeneity of the non-native speaker group, such a database may not be easily available. The above factors have severely restricted the spread of ASR technology in computer-aided language learning (CALL) [1].

Witt and Young [2] and Franco et al. [3] have described pronunciation scoring systems focused on measuring pronunciation quality of a non-native speaker at phone level. They used acoustic likelihood-based methods for automatic pronunciation assessment within the framework of a Hidden Markov Model speech recognition system, trained on native speech. Franco et al. [3] presented a paradigm for automatic assessment of pronunciation quality by machine and used human-expert ratings to validate the machine scores on Americans speaking French. Pronunciation scoring in [2] uses the recognizer in forced alignment mode (using the canonical transcription of the known read out speech) whereas [3] recognizes the utterance using free decoding mode. Pronunciation scoring applications aim at detecting and locating articulation errors. This cannot be achieved by using the canonical transcription alone for forced alignment. In the absence of training data corresponding to non-native speech in L2, phone recognition accuracy can be improved by using a mix of phone models derived from target language (L2), spoken by native speakers of L2 and native language (L1) phone models [1,4]. Further, decoding errors may be minimized by operating the recognizer in a constrained mode where the output phone sequence is restricted to one of a set of reasonable variations derived from the canonical utterance. Pronunciation variation modeling has been an important part of ASR research [5] and can be adapted for the pronunciation scoring task.

We present the design of a pronunciation scoring system for Indian learners of English. Indian speakers present a wide variety of accents depending on their native languages, geographical region and socio-economic background. We therefore consider the realistic scenario where a suitable Indian English speech database is not readily available. The only training databases available are native English and native Hindi speech. We evaluate different approaches to selecting acoustic models for the phone recognition system from the combination of the above L2 and L1 models. We discuss approaches to generate reasonable pronunciation variations of the canonical utterance. An evaluation of the pronunciation scoring system, developed using two popular speech recognition toolkits Sphinx-3 [6] and HTK3.4 [7], is carried out on a small set of speakers and sentences in terms of correct detection and likelihood scores. The results are discussed and suggestions for improvement are provided.

2. DATABASES FOR TRAIN AND TEST

Our chosen task is the pronunciation scoring of English as spoken by Indian speakers, making English the target language (L2). We use an available database of read English speech, the TIMIT corpus, as the native target language database. TIMIT, designed to provide speech data for the development and evaluation of ASR systems, contains 16 kHz sampled recordings of 630 speakers of eight major dialects of American English, each reading 10 phonetically rich sentences. Of the 10 sentences, 2 are common across all speakers. The two sentences have been designed specifically to incorporate phones that are relatively common in American English. The TIMIT database includes time-aligned word and phonetic transcriptions.

Our test data for the evaluation of the pronunciation scoring system comprises English sentences, selected from the TIMIT dataset, read out by 30 Indian college students (17 male, 13 female). The speakers were from across the country with different native tongues including Marathi, Telugu, Kannada, Tamil, Malayalam, Oriya, and Bengali. Each speaker was recorded in a quiet lab reading out two English sentences each, at 16 kHz sampling. The two sentences correspond to the TIMIT database common sentences and are shown in Table 1 with their phonetic transcriptions. These transcriptions were obtained from the annotations of a model Indian speaker of English and are considered canonical forms in the present study. The Indian English speech dataset is manually transcribed to obtain the reference or surface transcriptions to be used later for system evaluation. The phone set used for the transcription is the union of the TIMIT American English phone set and that of Hindi as obtained from the TIFR Hindi speech database described next.

Ideally, the acoustic models used in the ASR component of the pronunciation scoring system should be trained on Indian English speech for optimum performance. Such a database being unavailable, we consider the possibility of using an available Hindi speech corpus to compensate for phones not present in the L2 inventory that may be used by the non-native speaker. The TIFR Hindi speech database [8] comprises phonetically rich Hindi sentences uttered by 100 native speakers of Hindi. Each speaker utters 8 unique and 2 common sentences. The speech is recorded in quiet at 16 kHz sampling. The database includes word and phonetic transcriptions and is suitable for training phone level acoustic models for Hindi speech recognition. Our approach is similar in spirit to the use of a speaker-independent bilingual phone recognizer for pronunciation error detection [4].

Table 1. Evaluation TIMIT sentences SA1 and SA2 with bilingual phone transcriptions corresponding to a model Indian speaker of English

She	**had**	**your**	**dark**	**suit**
/S/I	/h/E/vbD/D	/y/O	/vbD/D/A/clk/k	/s/U/clT/T
in	**greasy**	**wash**	**water**	**all**
/i/n	/g/r/I/s/I	/w/O/S	/w/O/clT/T/er	/O/l/
year				
/y/er				

Don't	**ask**	**me**	**to**
vbD/D/O/N/clT/T	/A/s/clk/k	/m/I	/clT/T/u
carry	**an**	**oily**	**rag**
/clk/k/E/r/I	/E/n	/oy/l/I	/r/E/vbg/g
like	**that**		
/l/ay/vbg	/d/E/clT/T		

3. PRONUNCIATION SCORING SYSTEM

The pronunciation scoring system is designed to obtain separate articulation and prosodic scores for the learner's input utterance with respect to the canonical form stored in the system. To control for the effects of speech recognition errors, it is important to constrain the recognition process by limiting the output phone sequences based on some knowledge of non-native speaking errors. Accordingly the phone decoder is operated in forced alignment mode as shown in Figure 1 and the most likely outcome of the constrained set is considered the uttered phone sequence to be used for articulation and prosody scoring. The system blocks are described next.

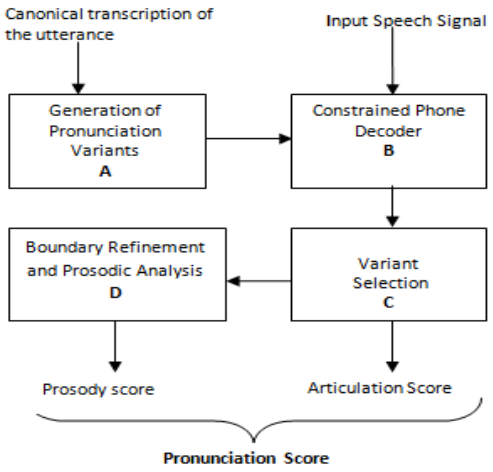

Figure 1. Pronunciation scoring system block diagram

A. Generation of pronunciation variants: The purpose is to obtain a set of highly probable variations given the orthographic/canonical form of the sentence provided to the learner to read out. To achieve this, it is necessary to take into account phonological variations arising due to the influence of the learner's L1. In the present set-up, we have learners from different regions of India and hence distinctly different L1s to whom different rules are applicable [9]. Since the scope of the present study is limited to the two sentences of Table 1, we manually constructed a pronunciation dictionary for the given words based on observations of the test non-native speakers together with known generalizations. The word pronunciations included phone substitutions, insertions and deletions that arise from phonotactic constraints in the native language [9]. Further, different combinations of word sequences were generated based on some simple rules of consistency in order to obtain a much larger set of plausible phone sequences for testing the system.

B. Constrained phone decoder: is an HMM-based phone recognizer. Our system is tested with two implementations of the phone recognizer, one based on HTK 3.4 and the other developed with Sphinx 3. Monophone, context-independent acoustic models are trained with speech from the selected language database (TIMIT American English and TIFR Hindi). The parameter settings for each system are provided in Table 2. The phone decoder is operated in forced alignment mode on each of the generated variants to obtain the phone segmentation boundaries along with likelihood scores of each phone and of the overall utterance.

C: Variant selection: The likelihood scores of the entire set of variants are ordered in decreasing value and the one with the highest likelihood is selected as the detected phone sequence.

This can be used to provide visual feedback to the learner in terms of phone substitutions, insertions and deletions with respect to the canonical transcription of the corresponding sentence. Additionally, an articulation score can be derived from the detected phone errors and also possibly phone likelihoods.

D. Boundary refinement and prosodic analysis: where pitch contour and relative durations are matched against corresponding syllables of a model speaker. This block is yet to be implemented.

Table 2. Feature Configuration parameters for HTK 3.4 and Sphinx3

Parameter		HTK3.4	Sphinx3
# of HMM states used for training		3	3
# of Gaussian Mixtures	TIFR models	8	16
	TIMIT models	24	16
Pruning beam width parameter		0.0	1e-64
Frame length		25 ms	25.6 ms
Frame rate		100/s	100/s
Mel Filters Number		26	40
Pre-emphasis		0.97	0.97
Energy Normalization		True	True

4. EXPERIMENTS AND EVALUATION

The focus of this work is to investigate the effect of the choice of phone inventory and training data for the acoustic phone models on the performance of the pronunciation scoring system. The performance is evaluated in terms of the non-native speech recognition accuracy as explained later. Experiments are carried out with the following decoder configurations.

A. 47 class TIMIT: L2 native phone models only are used. The phone inventory comprises 47 American English phones after suitable folding over of the original distinct 63 phones and the acoustic models are trained on the TIMIT train set.

B.52 class: the previous L2 phone models are augmented by Hindi (L1) phone models. A union of TIMIT and TIFR Hindi database phone inventories is created. The 47 American English phones in this inventory are represented by acoustic models trained on TIMIT data. The 5 phones of Hindi not included in the American English phone inventory are added via acoustic phone models trained on the TIFR Hindi database.

C.48 class: Native-language (L1) phone models augmented by L2 phone models. Similar to B. above, but now the majority of the phone models are trained on the TIFR Hindi database with only the non-overlapping phones obtained by training on TIMIT.

The above choices represent three distinct methods of creating a phone model set suitable for the Indian English speech decoder. In each case, the canonical forms of the sentences to be read, their generated variants as well as the surface transcriptions of the non-native speakers are mapped to the corresponding phone inventory. Forced-alignment of each variant with the non-native speaker's acoustic feature sequence is carried out to obtain a likelihood score for the variant.

The evaluation is carried out on the top-ranked variant phone sequences and associated likelihoods using three different measures. The measures are designed to capture the suitability of the system for reliable scoring of non-native pronunciation. These are (i) the number of instances in which the surface transcription is within the top N decoded sequences in terms of likelihood score, (ii) the edit distance between the most likely phone sequence and the surface transcription in terms of %correct and %accuracy, and (iii) normalized likelihood error given by the difference between the likelihood scores of the best and surface phone sequences divided by the surface likelihood score. A value of "0" for this measure indicates the best achievable performance.

5. RESULTS AND DISCUSSION

Table 3 shows the free decoder performances on the non-native speech test set for HTK 3.4 and Sphinx 3 phone recognizers with acoustic models corresponding to each category described in Sec. 4. The word insertion penalty was optimized to ensure that the phone insertion and deletion rates were equal. The null grammar language model was used. We note that the overall phone recognition accuracies are poor with the HTK 3.4 system doing better than the Sphinx 3 system. A good language model could have served to improve performance to some extent. However, on the flip side, this could obscure the very pronunciation errors that the system is expected to detect.

The 48-class acoustic models provide the best performance in both systems indicating that the non-native (Indian English in our case) speech is significantly better matched with models that are predominantly drawn from native Hindi speech rather than from the target language (English) native speech.

Table 3. Phone recognition accuracies by the HTK 3.4 and Sphinx 3 decoders

Decoder Models	WIP for HTK-3.4	HTK-3.4 %Acc	WIP for Sphinx3	Sphinx3 %Acc
52-class	-10.3	26.68	5e11	34.42
48-class	-8.15	46.12	5e12	36.32
47-class	-10.5	29.05	7e11	33.78

The performance of the HTK 3.4 and Sphinx 3 decoders operating in the forced alignment mode as depicted in the pronunciation scoring system of Fig. 1, are shown in Tables 4 and 5. The phone sequences for forced alignment were generated by the method discussed in Sec. 3 A, by applying rules to each canonical phone transcription. In the Tables 4 and 5 are shown the number of instances (out of 30 utterances for each sentence) that rank in the top 1 and top 5 with respective decoder likelihood scores. The surface transcription does not always rank 1 due to the limitations of the phone recognizer. HMM decoding is optimized for classification and not boundary alignment. Hence there is a possibility that the recognizer would assign higher likelihood score to a variant that is close enough to the surface transcription, i.e. the transcription with the highest score and the surface transcription may differ by only one vowel which may cause the surface transcription to have a lower score. From the top N counts, we observe that the 48-class phone models work best in terms of correctly selecting the transcription corresponding to the ground-truth (surface) transcription. In cases where the most likely transcription does not match the surface transcription, the edit distance is non-zero. The average %correct and %accuracy achieved across the test sentences are shown in Tables 4 and 5. The trend with the 48-class model providing the best match is consistent with its decoding accuracy as well as ranking performance. This may be explained by the fact that speakers have used the phones from their native tongue rather than the target language phones. The two diphthongs (/ay and /oy) in SA2 were recognized with 100% accuracy despite using the models for these labels from native English trained set. Hence we can conclude that although an Indian speaker of English substitutes English phones with phones from his native tongue closest to the corresponding English phone,

nevertheless uses the correct diphthong if no corresponding native phone is available.

Table 4. Forced Alignment Decoding by HTK 3.4

Decoder models	# of Unique variants	METHOD I: Reference transcription in		METHOD II: %Corr	METHOD II: %Acc
		Top1	Top5		
SA1					
52-class	1263	6	9	81.83	80.00
48-class	763	21	24	96.20	94.60
47-class	636	5	7	82.43	79.39
SA2					
52-class	1026	7	13	88.58	86.96
48-class	1026	16	20	92.75	92.24
47-class	1026	6	11	87.98	85.76

Table 5. Forced Alignment Decoding by Sphinx3

Decoder models	# of Unique Variants	METHOD I: Reference transcription in		METHOD II: %Corr	METHOD II: %Acc
		Top1	Top5		
SA1					
52-class	1263	1	6	82.51	78.40
48-class	763	12	17	92.24	89.13
47-class	636	2	6	83.80	80.15
SA2					
52-class	1026	5	10	85.85	83.80
48-class	1026	7	9	89.43	87.72
47-class	1026	5	9	88.75	86.36

Finally, the Figure 2 shows the distribution of the likelihood scores across the 60 utterances for each phone set giving us a more complete picture of the performance than is indicated by the top N indices alone. We observe that the 48-phone class has average likelihood error closest to zero of the three phone sets. The likelihood errors are more spread in the case of the Sphinx 3 decoder. There are a number of outliers with high likelihood errors observed in nearly every configuration. A closer study of the outliers revealed the following characteristic decoding errors:

- Burst followed by a fricative or vice versa, will not be recognized if the burst is not strong.
- Burst followed by a nasal will not be recognized if the burst is not strong.
- Poor recognition for voice bar when followed or preceded by a nasal.
- Poor recognition of rhotic /r when it is preceded or followed by a vowel.
- /hv (voiced) preceded by vowel and followed by fricative is not recognized correctly
- Poor recognition for shwa followed or preceded by /A or /a
- Poor recognition in case variant has more phone deletions when compared to reference transcription.

Some of the above errors can be attributed to the mismatch of the native language model phones (trained on native Hindi speech from the TIFR database) with the true native language of the Indian English speaker (drawn from various regions of the country). This demonstrates that a multilingual (rather than bilingual) phone set may be necessary for Indian English pronunciation scoring.

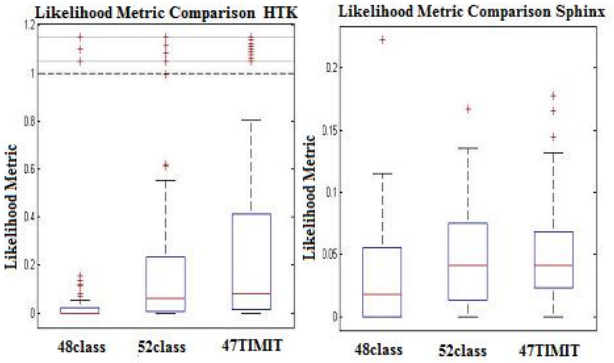

Figure 2. Box plot of Likelihood error metric for HTK3.4 and Sphinx3

6. CONCLUSIONS

Phone recognition errors pose a serious obstacle to the deployment of ASR in a computer-aided language learning application. Using forced alignment for phone decoding with a suitable set of pronunciation variations can compensate for these inherent performance shortcomings. In the case of Indian English learners, a bilingual phone recognizer with phone inventory and acoustic models drawn from native Hindi speech supplemented by American English data trained models for the missing phones of English worked relatively well. Future work should take into considerations the observed specific errors of the decoder in order to design better features for recognition. The procedure for generation of pronunciation variations needs to be generalized and automated in order to be applicable to arbitrary sentences. It is not clear why the performances of the HTK 3.4 and Sphinx 3 decoders were so different. An understanding of this should help to tune both decoders towards better performance.

7. REFERENCES

[1] Strik, H., Neri, A., and Cucchiarini, C. 2008. Speech technology for language tutoring. In *Proceedings of LangTech* (Rome, Italy,February 28-29, 2008).

[2] Witt, S., and Young, S. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication.* Vol. 30, pp. 95-108, 2000.

[3] Franco, H., et al. 2000. Automatic scoring of pronunciation quality. *Speech Communication.* Vol. 30, pp. 83-93, 2000.

[4] Kawai, G., Hirose, K. 1998. A method for measuring the intelligibility and non nativeness of phone quality in foreign language pronunciation training. In *Proceedings of ICSLP-98* (Sydney, Australia, November 30- December 04,1998) .pp. 1823-1826.

[5] Goronzy, S., Rapp ,S., Kompe, R. 2004. Generating non-native pronunciation variants for lexicon adaption. *Speech Communication.* Vol. 42, pp. 109-123, 2004.

[6] Lee, K.F. 1998. *Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system.* Ph.D. dissertation, Comput. Sci. Dep., Carnegie Mellon University.

[7] Young, S., et al. 2006. *The HTK Book v3.* Cambridge University, 2006.

[8] Samudravijaya, K., Rawat, K.D., and Rao, P.V.S. 1998. Design of Phonetically Rich Sentences for Hindi Speech Database. *J. Ac. Soc. Ind.* Vol. XXVI, December 1998, pp. 466-471.

[9] Gargesh, R. 2004. *Indian English: Phonology.* A handbook of varieties of English: a multimedia reference tool, Volume 1.